Housing Prices Prediction

BIA 652
Fall 2018
Alex Caruso and Alicia Kohl

copyright © 2008-07-10 : sean dreilinger : http://durak.org/sean/

Background Info

Housing dataset from Kaggle for home prices in King County in Washington State, which includes Seattle, from May 2014-2015. There were approximately 22,000 observations.

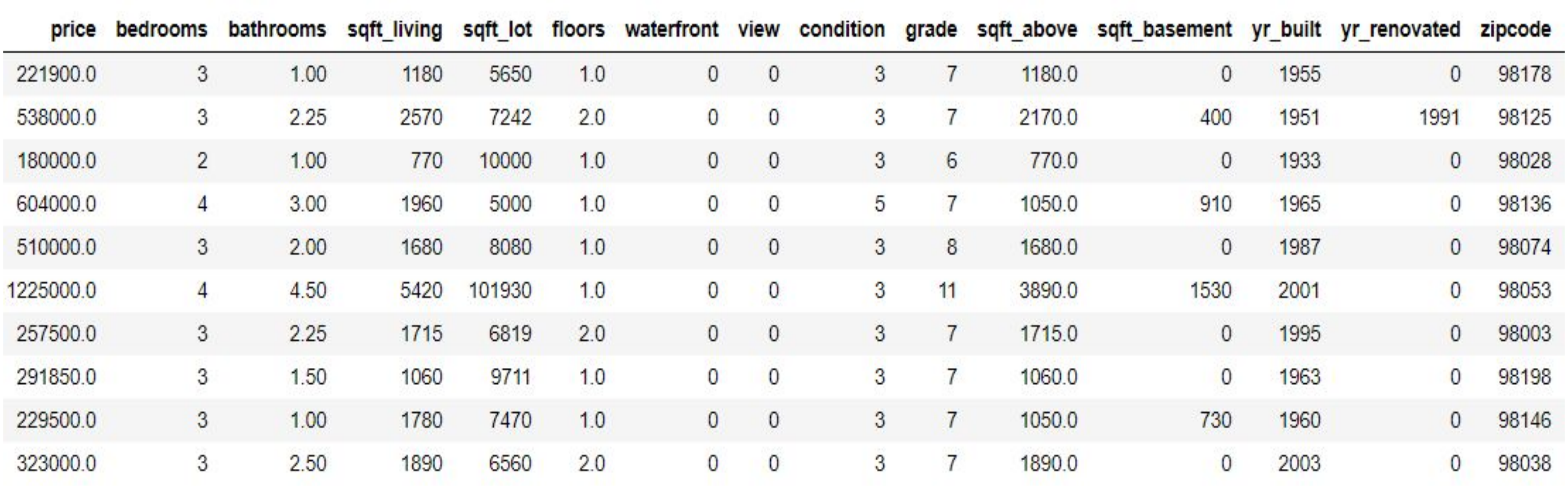
The data will be used to accurately predict housing prices and the weight different attributes such as number of bedrooms, type of view, etc. have on the price of homes in the given area.

Problem Statement

- **Problem Statement:** Accurately predict home prices based on their characteristics and dimensions
- **Application:** Services like Zillow, Real Estate Firms, and Independent Home Seller or Buyers can benefit from strong predictive models for home pricing
- **Potential Business Impacts:**
    - Quicker home sales
    - Fair deals for buyers and sellers due to more accurate pricing
    - Determination of most important features in home value
        - This helps renovators / house flippers optimize their spending

Dataset

Raw Data Snapshot

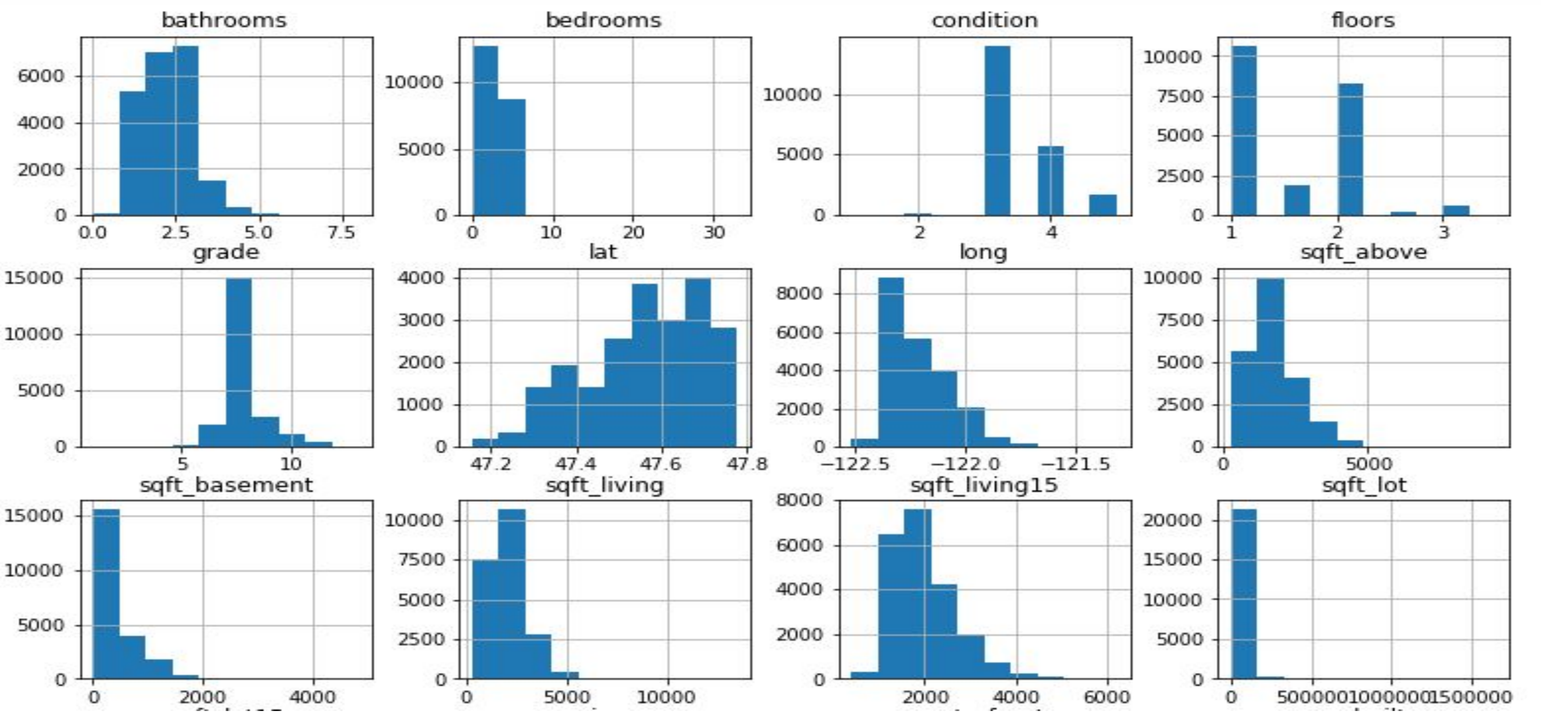


price	bedrooms	bathrooms	sqft_living	sqft_lot	floors	waterfront	view	condition	grade	sqft_above	sqft_basement	yr_built	yr_renovated	zipcode
221900.0	3	1.00	1180	5650	1.0	0	0	3	7	1180.0	0	1955	0	98178
538000.0	3	2.25	2570	7242	2.0	0	0	3	7	2170.0	400	1951	1991	98125
180000.0	2	1.00	770	10000	1.0	0	0	3	6	770.0	0	1933	0	98028
604000.0	4	3.00	1960	5000	1.0	0	0	5	7	1050.0	910	1965	0	98136
510000.0	3	2.00	1680	8080	1.0	0	0	3	8	1680.0	0	1987	0	98074
1225000.0	4	4.50	5420	101930	1.0	0	0	3	11	3890.0	1530	2001	0	98053
257500.0	3	2.25	1715	6819	2.0	0	0	3	7	1715.0	0	1995	0	98003
291850.0	3	1.50	1060	9711	1.0	0	0	3	7	1060.0	0	1963	0	98198
229500.0	3	1.00	1780	7470	1.0	0	0	3	7	1050.0	730	1960	0	98146
323000.0	3	2.50	1890	6560	2.0	0	0	3	7	1890.0	0	2003	0	98038

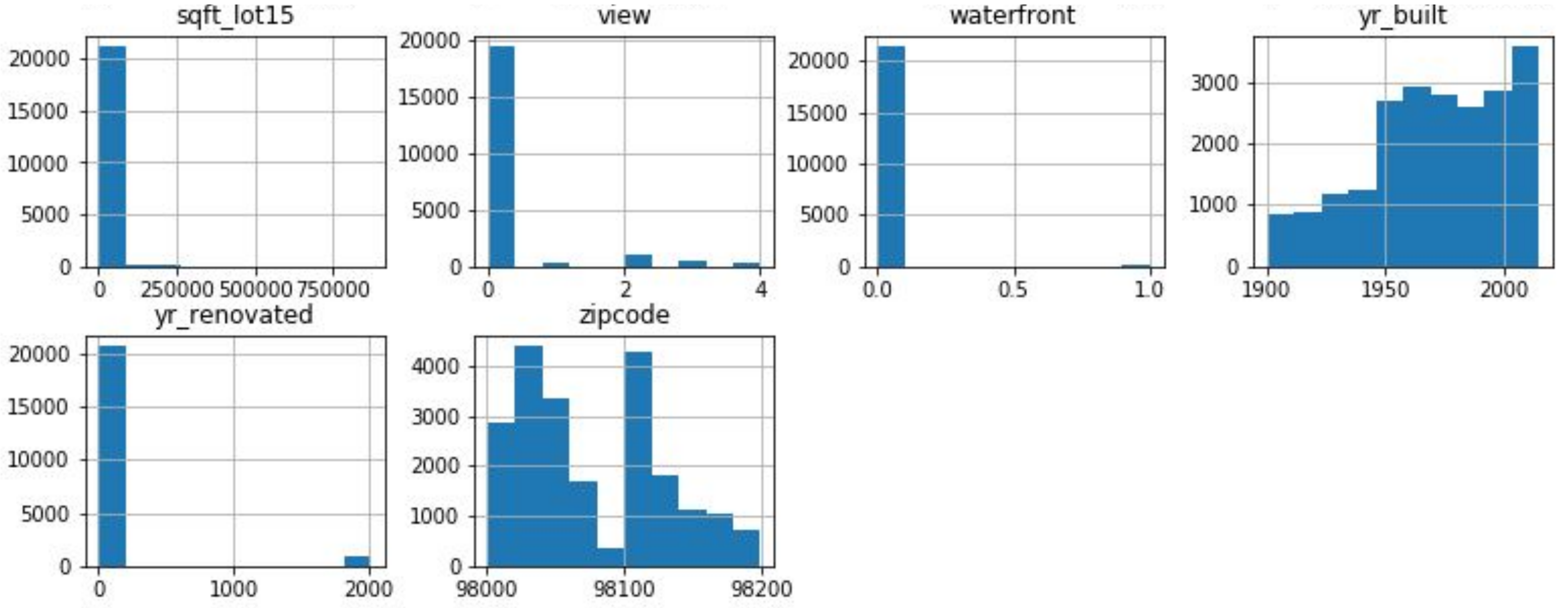
Descriptive Statistics

- 21,613 observations
- 21 attributes/features
- Average Home Price: $540,000
- Median Home Price: $450,000
- Range:
    - Min: $75,000
    - Max: $7,700,000
- Standard Deviation of Price: $367,000

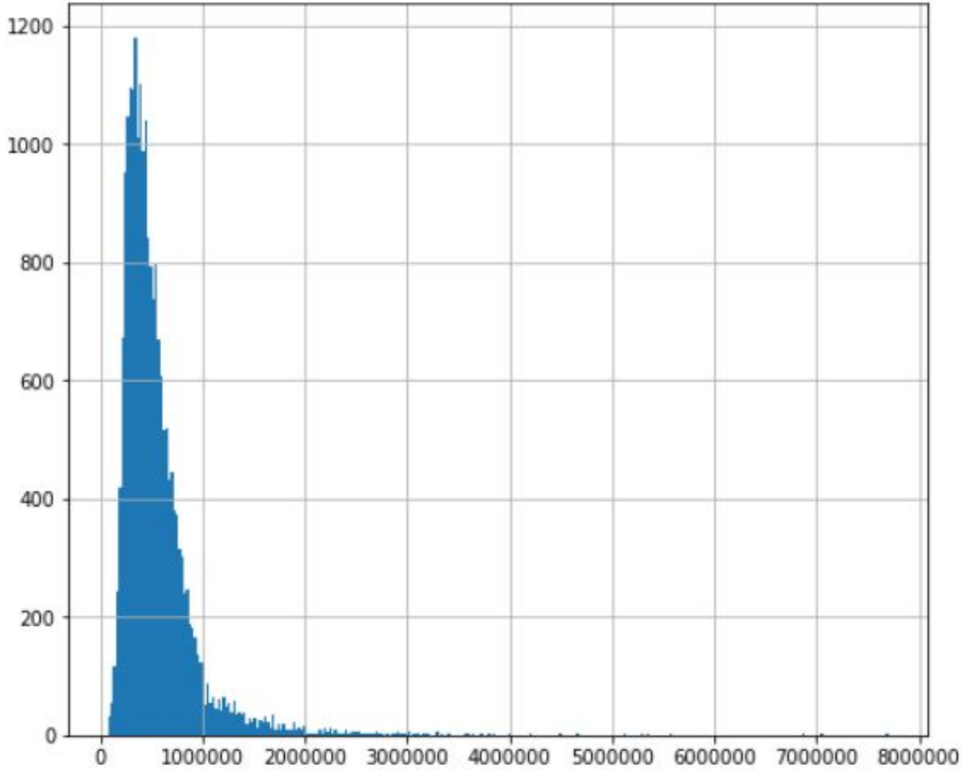
Independent Variable Distributions



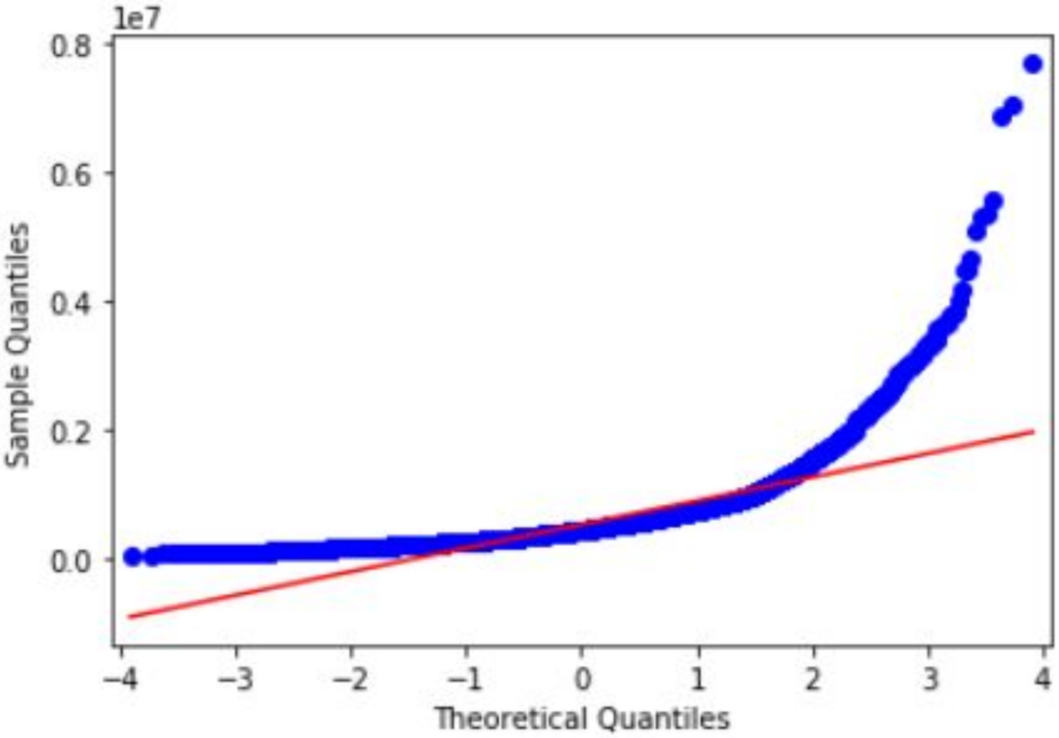
Independent Variable Distributions



Distribution of Price Variable

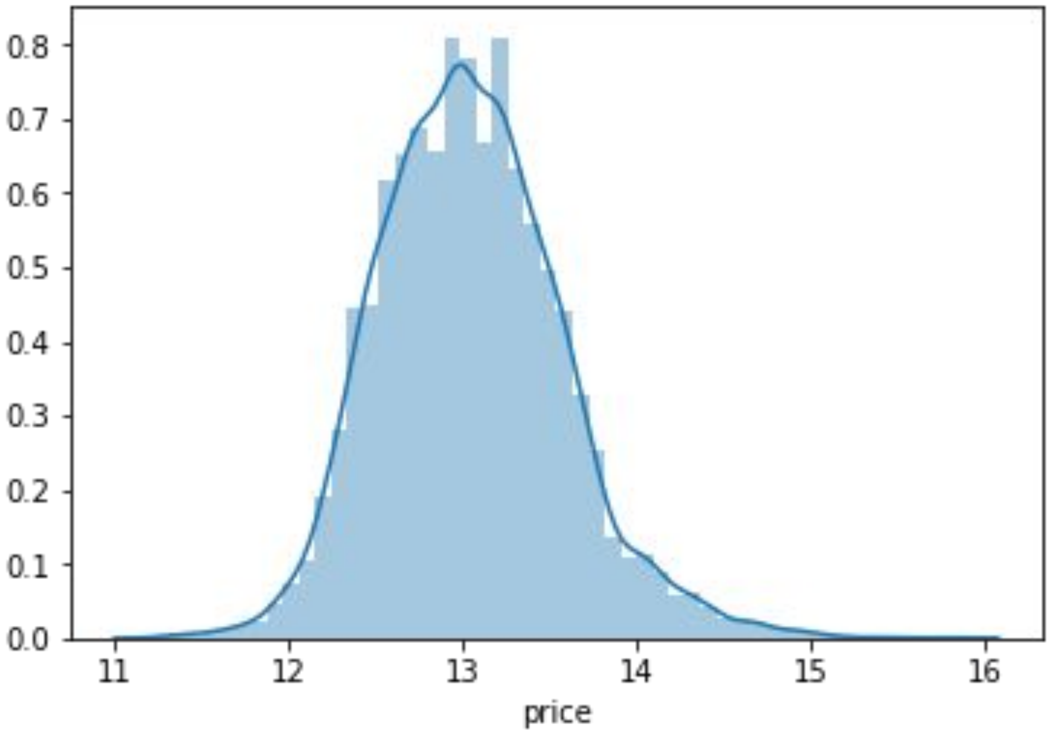


Price QQ Plot

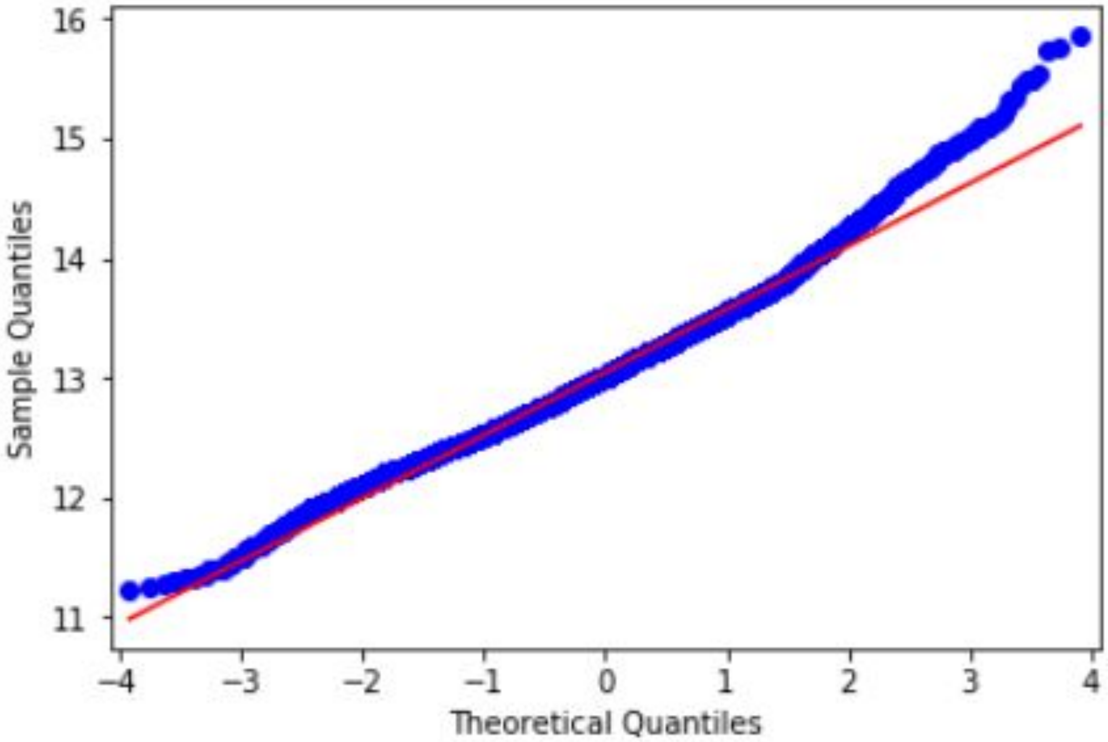


Logistic Transformation

Price Distribution following Logistic Transformation

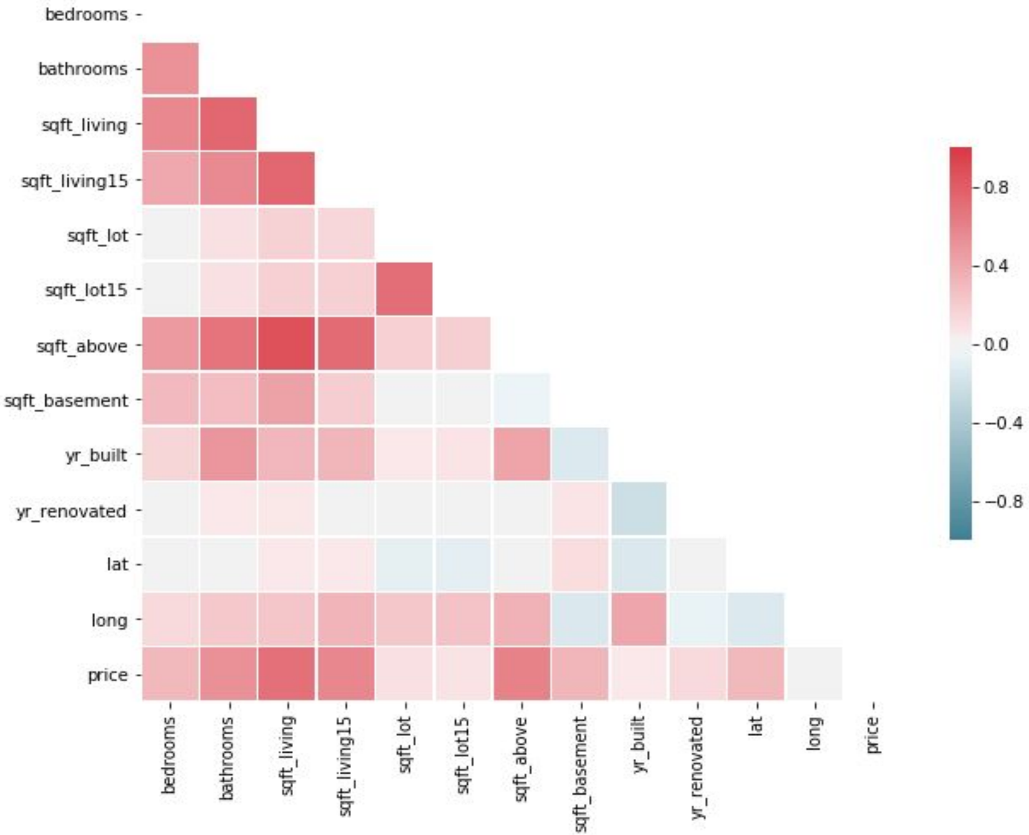


QQ Plot following Logistic Transformation

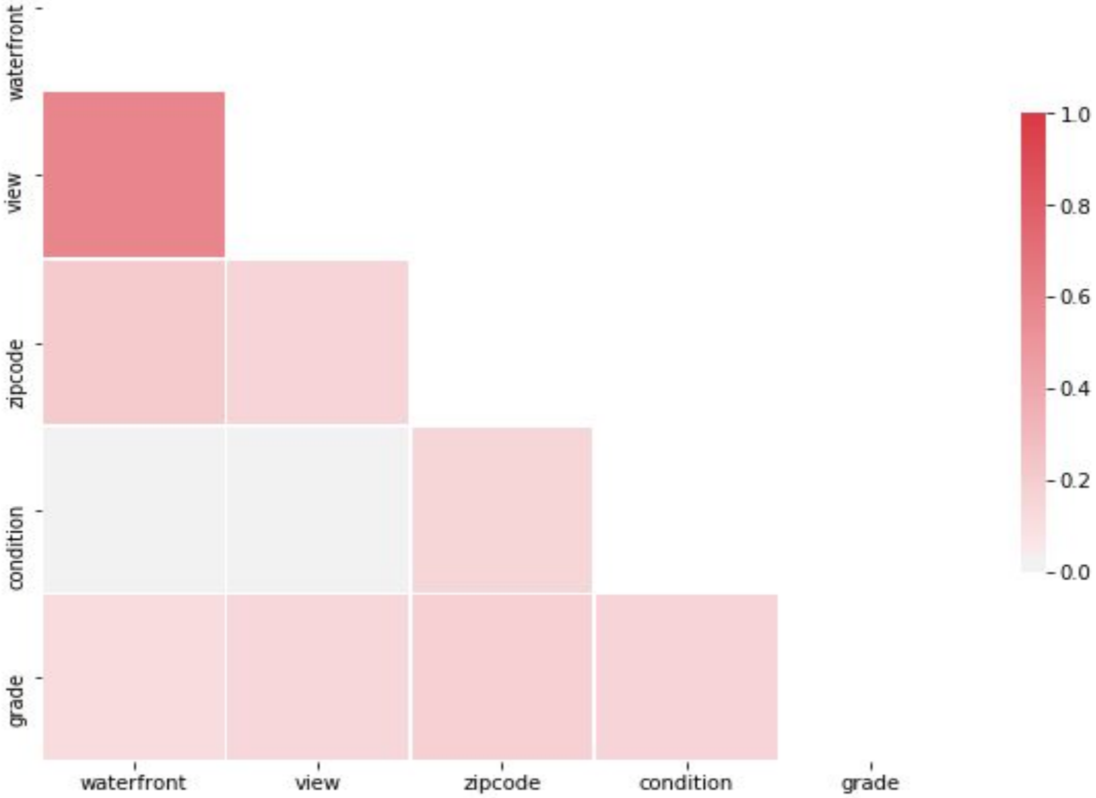


Correlation Analysis

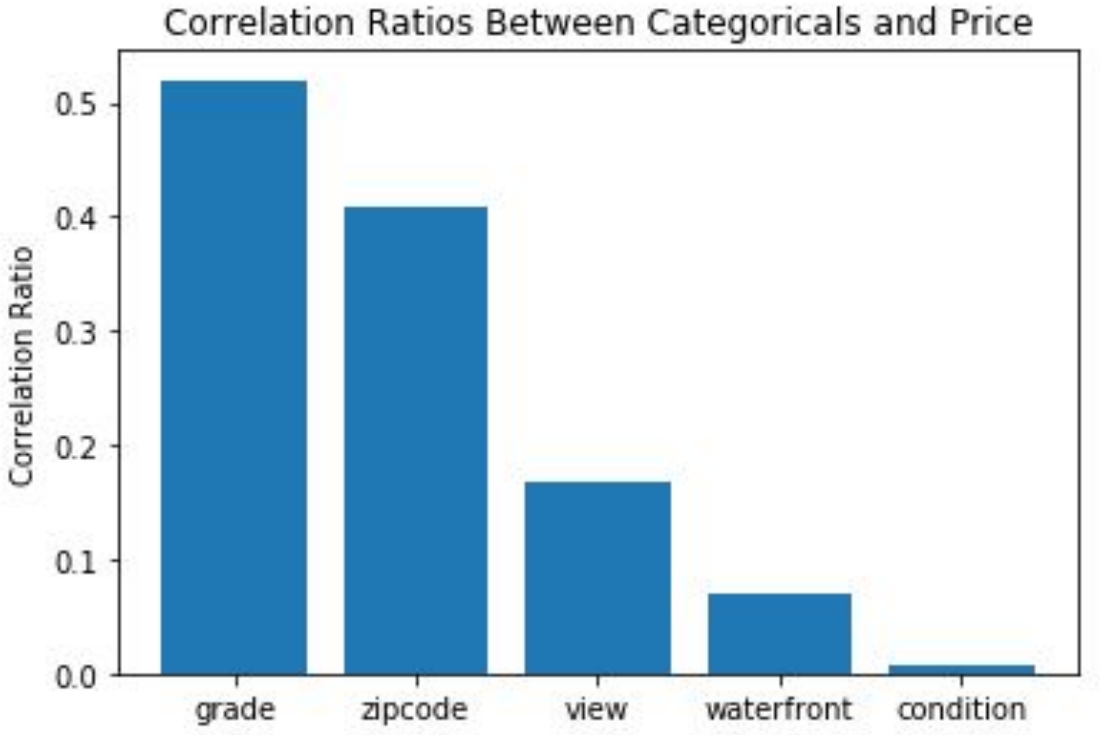
Continuous Variable Correlations



Categorical Variable Correlations



Correlation Ratios

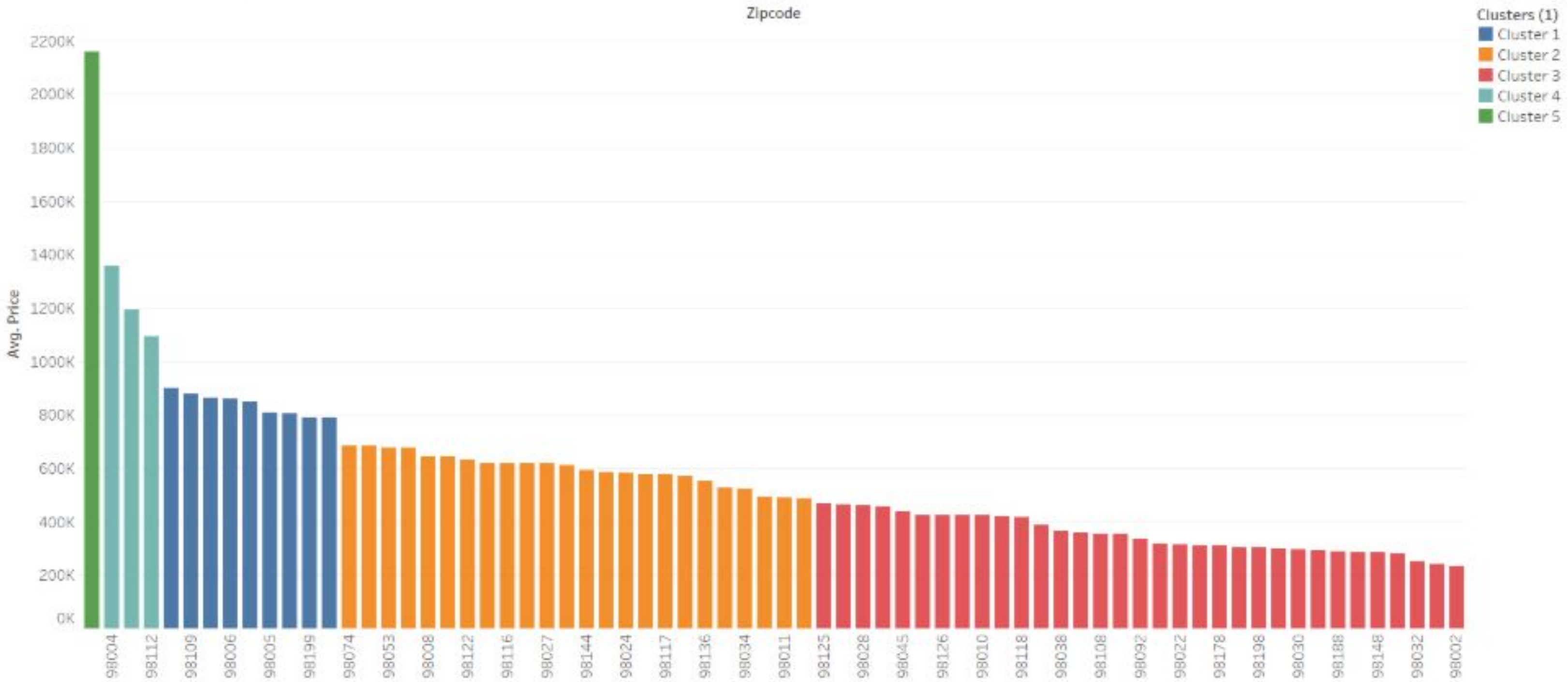


Data Preprocessing

Key Steps

- Filled null values with the avg for that column
- Created binary dummy variables for **year_renovated**, **view**, and **waterfront**
    - One-hot encoding: n-1 dummy columns where n is the number of unique values for the feature
        - Dropped one column from each ‘set’ of binary columns to reduce multicollinearity
    - Removal of **view_4** due to 0.6 correlation with waterfront_1
- Removal of **sqft_above, sqft_living15, and sqft_lot15** variables due to high correlations
- Normalized continuous variables for PCA
- Clustered zip codes by average home prices

Zipcode Clustering (Tableau)



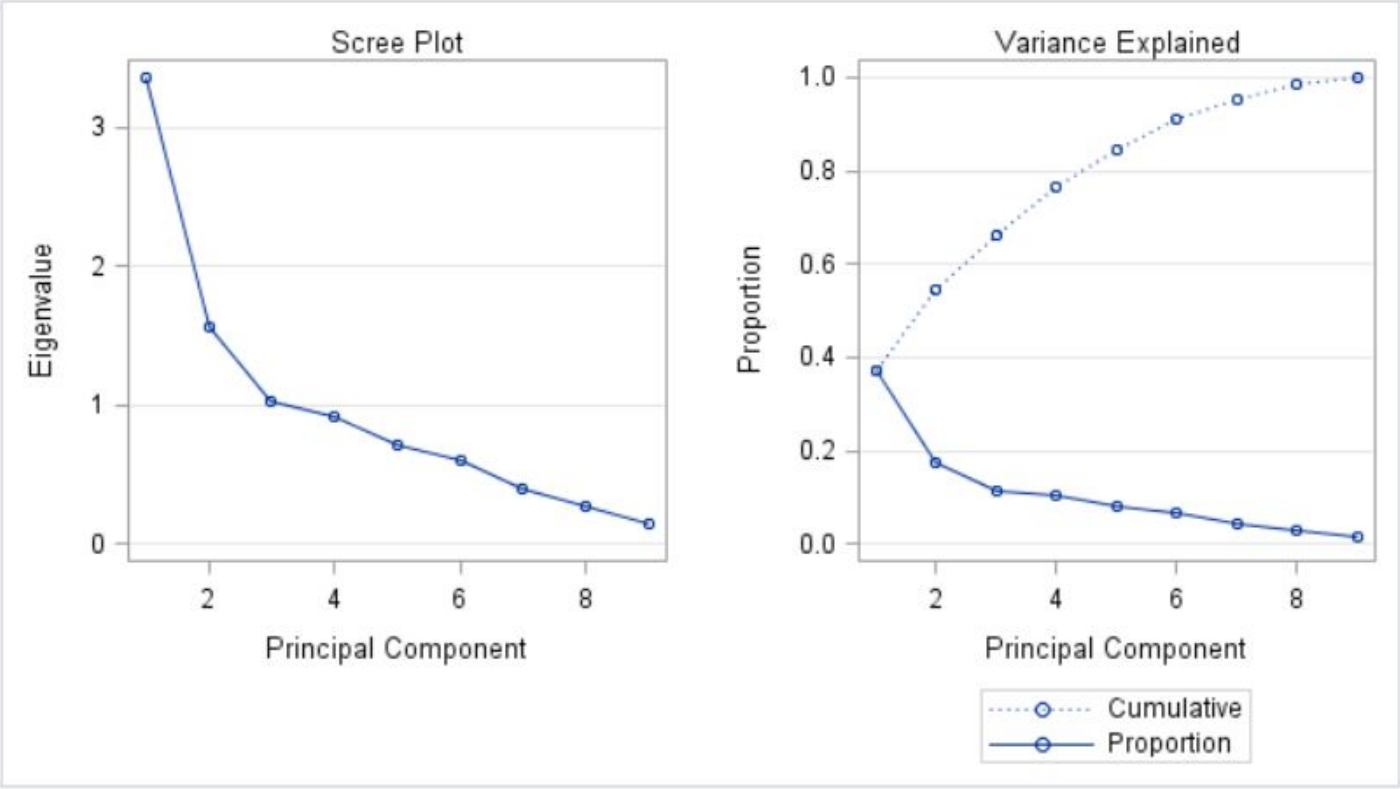
Zipcode Clusters by Avg Price

Principal Component Analysis

Overview of PCA

- Created 9 principal components from Z normalized continuous variables
    - Zipcode_cluster, bedrooms, bathrooms, sqft_living, sqft_lot, sqft_basement, yr_built, condition, grade
- Did not implement PCA for categorical variables
    - Not appropriate for one-hot encoded nominal features
- Resulting principal components were uncorrelated with categorical variables
- PCA was used to make features independent - not for dimensionality reduction

Screen Plot & Variance Explained



Multiple Linear Regression Model

Analysis of Influential Points

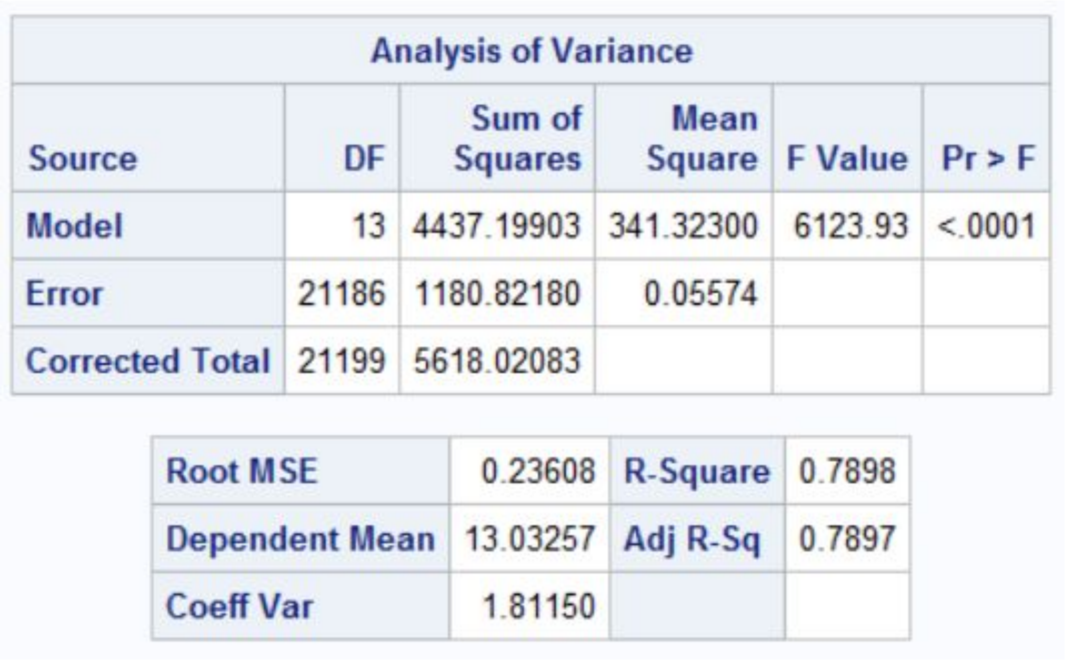
- After an initial regression (with all principal components and categorical variables), influential points were identified and removed from the dataset
- Cutoff formulas used:
  - Cook's D: 4/n
  - Dffits: [2*sqrt(p+1)] / [sqrt(n-p-1)]
  - Leverage: (2*p)/n
  - p=# of parameters
  - n=# of points

- # of high leverage points: 2462
- # of high Cook' D points: 1219
- # of high Dffits points: 593

- # of observations that exceed all 3 cutoffs: 413
  - These observations were removed from the dataset
  - After removing these observations, the final regression model was trained on the remaining data

R Squared, Model and Beta Significance

- Model is statistically significant with p-value <.0001
- Principal components 1-4 and 6-9 and the categorical variables are significant at the .05 level.
- Prin5 was statistically insignificant with a p-value of .37.

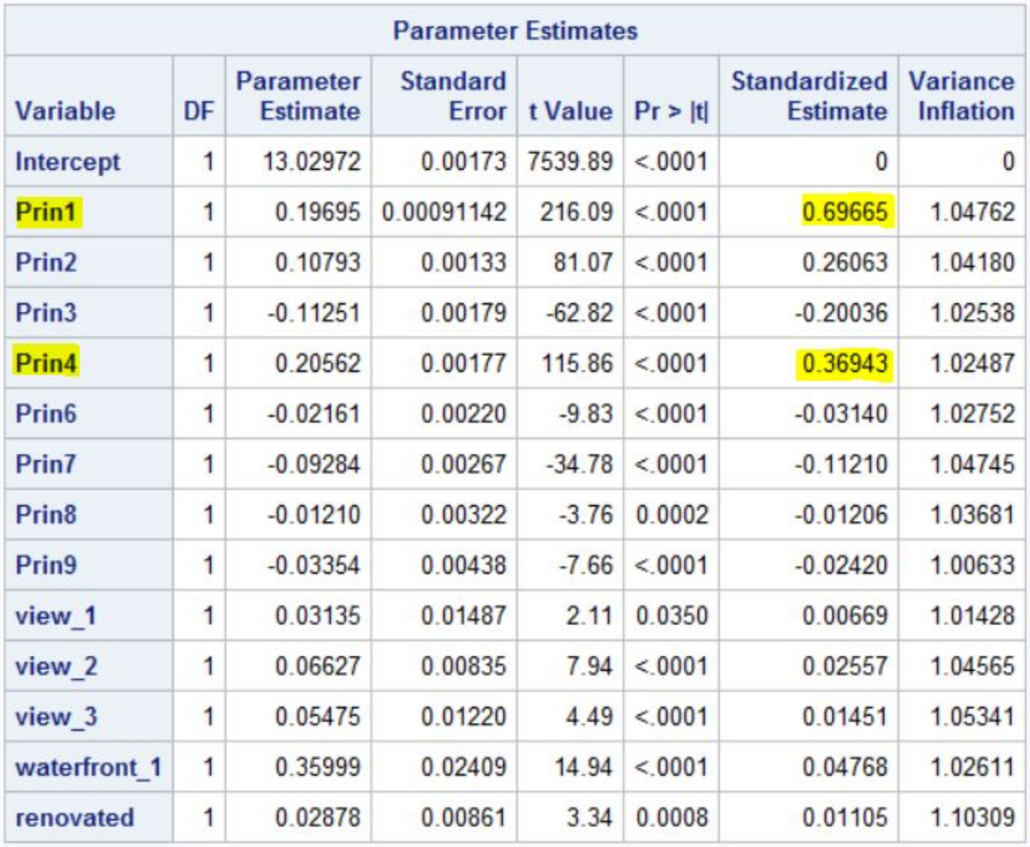


Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	13	4437.19903	341.32300	6123.93	<.0001
Error	21186	1180.82180	0.05574		
Corrected Total	21199	5618.02083			

Root MSE	0.23608	R-Square	0.7898
Dependent Mean	13.03257	Adj R-Sq	0.7897
Coeff Var	1.81150		

Model Betas & Variance Inflation

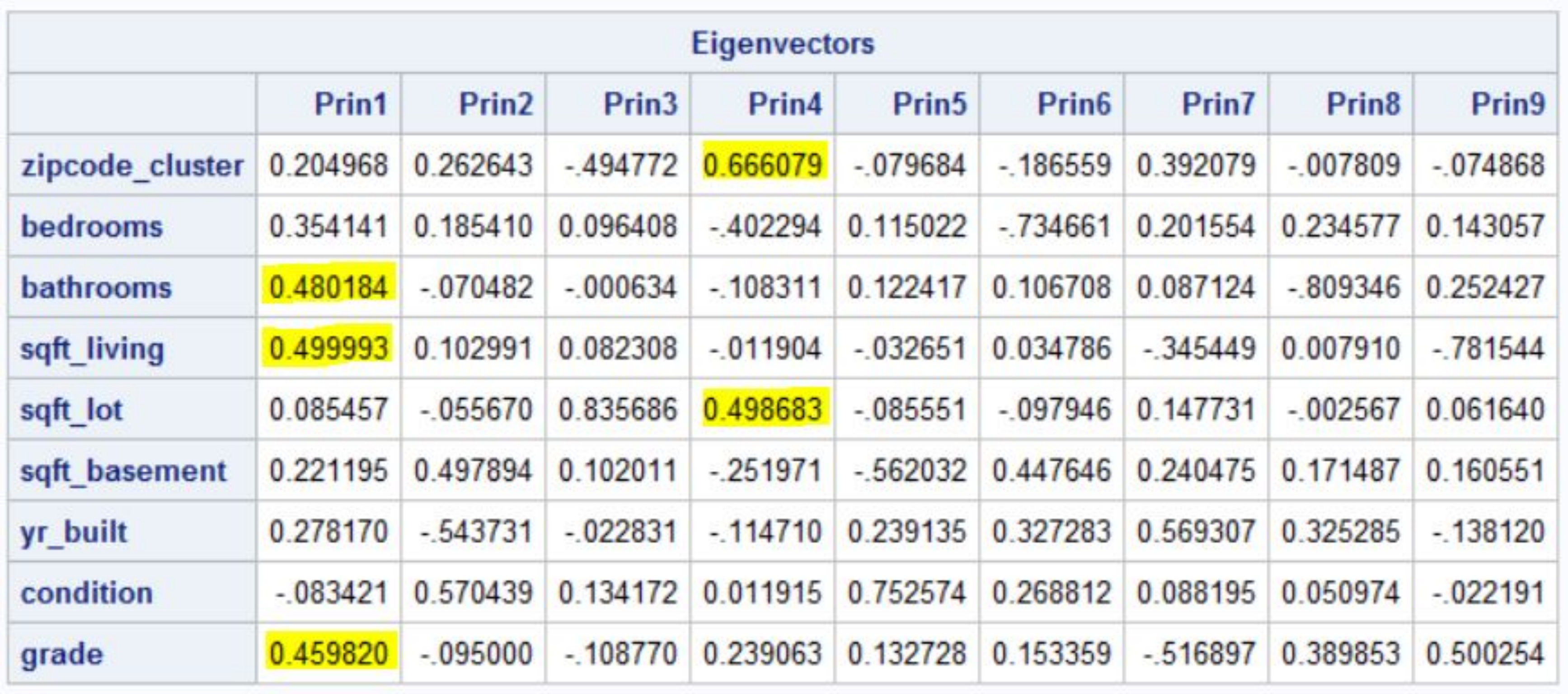


Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Standardized Estimate	Variance Inflation
Intercept	1	13.02972	0.00173	7539.89	<.0001	0	0
Prin1	1	0.19695	0.00091142	216.09	<.0001	0.69665	1.04762
Prin2	1	0.10793	0.00133	81.07	<.0001	0.26063	1.04180
Prin3	1	-0.11251	0.00179	-62.82	<.0001	-0.20036	1.02538
Prin4	1	0.20562	0.00177	115.86	<.0001	0.36943	1.02487
Prin6	1	-0.02161	0.00220	-9.83	<.0001	-0.03140	1.02752
Prin7	1	-0.09284	0.00267	-34.78	<.0001	-0.11210	1.04745
Prin8	1	-0.01210	0.00322	-3.76	0.0002	-0.01206	1.03681
Prin9	1	-0.03354	0.00438	-7.66	<.0001	-0.02420	1.00633
view_1	1	0.03135	0.01487	2.11	0.0350	0.00669	1.01428
view_2	1	0.06627	0.00835	7.94	<.0001	0.02557	1.04565
view_3	1	0.05475	0.01220	4.49	<.0001	0.01451	1.05341
waterfront_1	1	0.35999	0.02409	14.94	<.0001	0.04768	1.02611
renovated	1	0.02878	0.00861	3.34	0.0008	0.01105	1.10309

- Before PCA, some VIF values were between 5-10.
- After PCA, all are about 1.
- **Principal 1 and Principal 4 are the most important predictors**

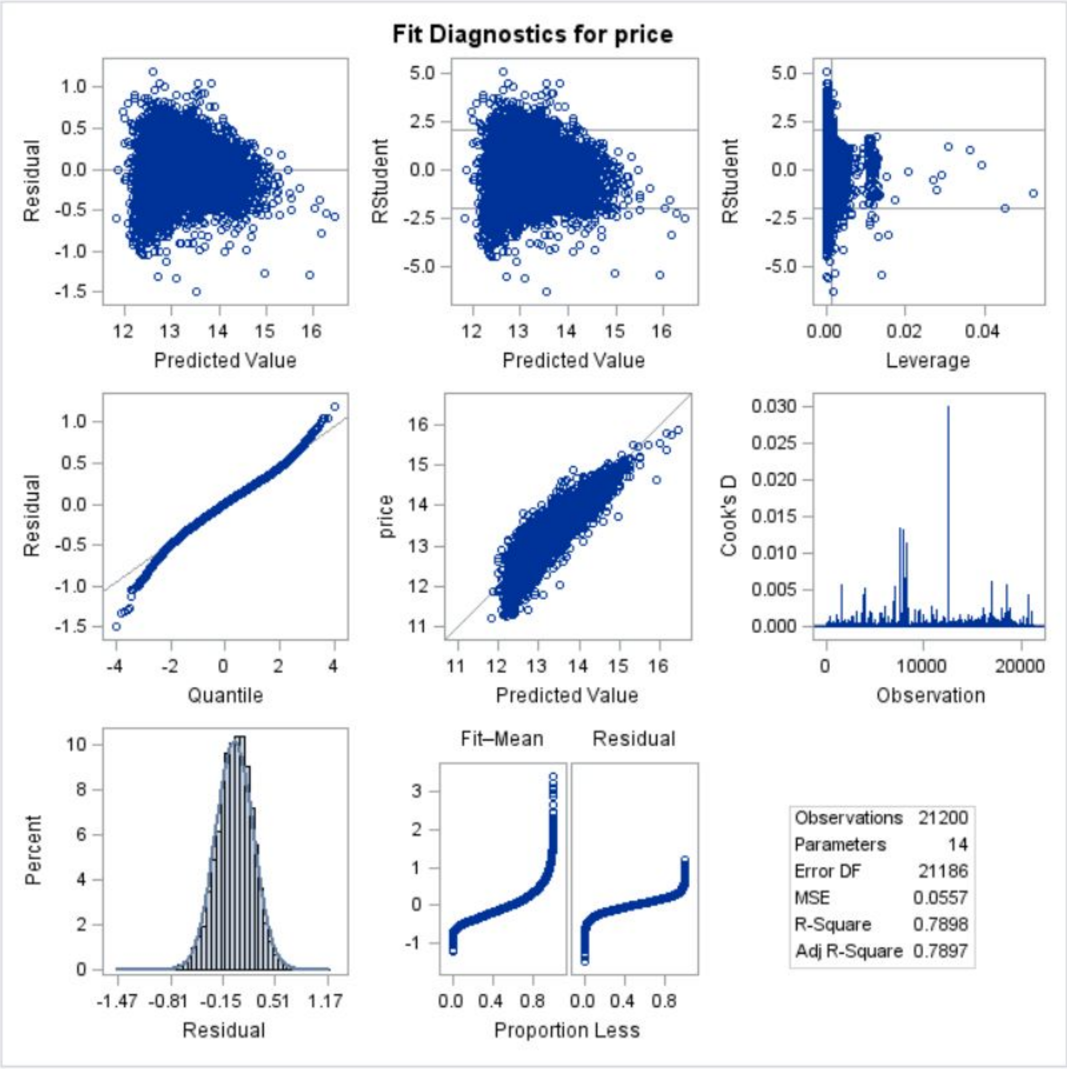
Analysis of Eigenvectors



Eigenvectors									
	Prin1	Prin2	Prin3	Prin4	Prin5	Prin6	Prin7	Prin8	Prin9
zipcode_cluster	0.204968	0.262643	-.494772	0.666079	-.079684	-.186559	0.392079	-.007809	-.074868
bedrooms	0.354141	0.185410	0.096408	-.402294	0.115022	-.734661	0.201554	0.234577	0.143057
bathrooms	0.480184	-.070482	-.000634	-.108311	0.122417	0.106708	0.087124	-.809346	0.252427
sqft_living	0.499993	0.102991	0.082308	-.011904	-.032651	0.034786	-.345449	0.007910	-.781544
sqft_lot	0.085457	-.055670	0.835686	0.498683	-.085551	-.097946	0.147731	-.002567	0.061640
sqft_basement	0.221195	0.497894	0.102011	-.251971	-.562032	0.447646	0.240475	0.171487	0.160551
yr_built	0.278170	-.543731	-.022831	-.114710	0.239135	0.327283	0.569307	0.325285	-.138120
condition	-.083421	0.570439	0.134172	0.011915	0.752574	0.268812	0.088195	0.050974	-.022191
grade	0.459820	-.095000	-.108770	0.239063	0.132728	0.153359	-.516897	0.389853	0.500254

Fit Diagnostics

- Fit diagnostics are for **log price**
- Residuals approximate a normal distribution
- R squared: 0.79
- Durbin-Watson of 1.97 so no evidence of autocorrelation among independent variables



Observations	21200
Parameters	14
Error DF	21186
MSE	0.0557
R-Square	0.7898
Adj R-Square	0.7897

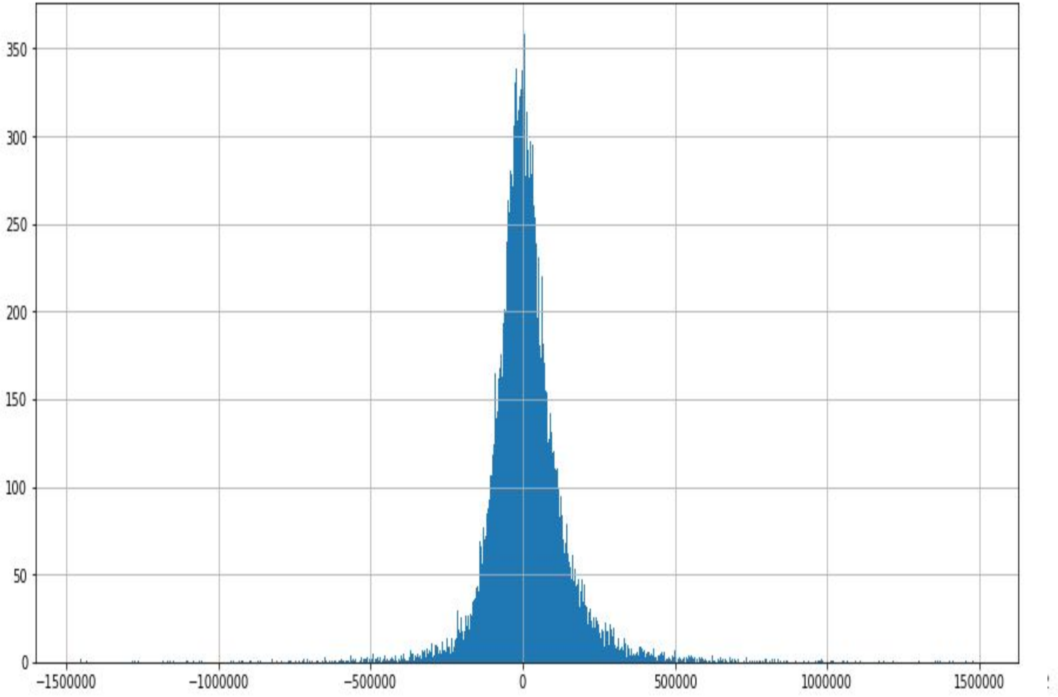
Predictions & Model Evaluation

Transforming Back to Original Scale & Predicting

- Take exponential of the log values for prices
    - Scaled predictions = e^(ln(price))
- R^2 based on original scale is 0.79
    - Some accuracy is lost when transforming predictions back to original scale
- Mean Absolute Error is $96,050
- Median Absolute Error is $65,520

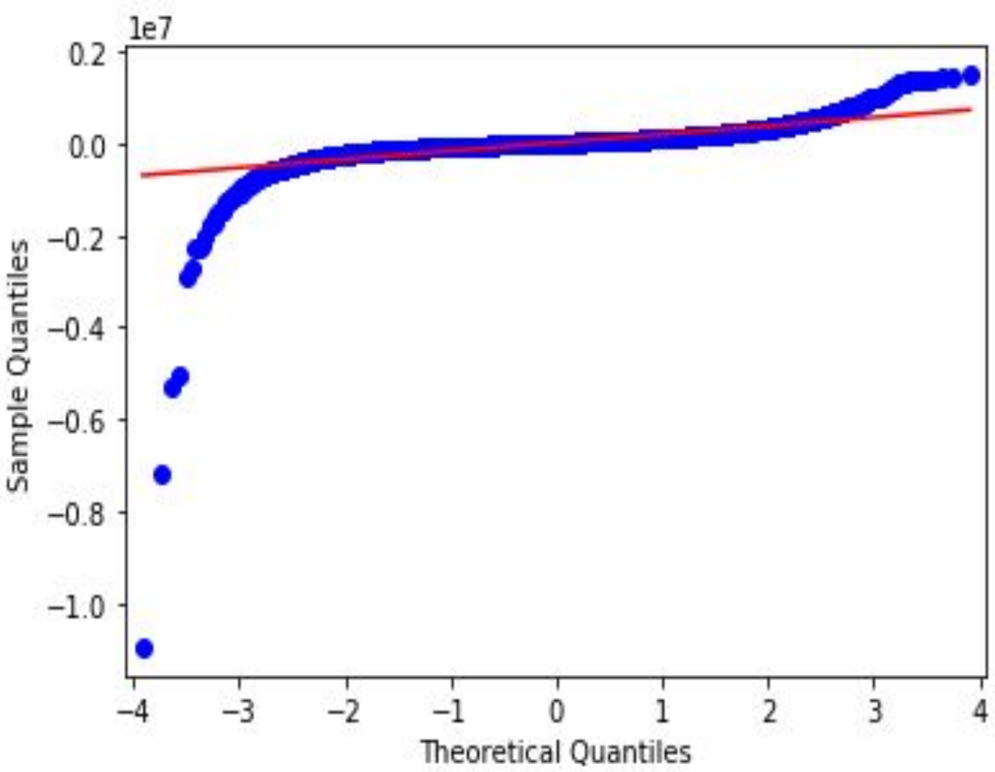
Residuals for Scaled Predictions

- Residual distribution is skewed due to outliers
- After filtering outlier residuals from the graph, the approximate normal distribution becomes clear



Residual and QQ Plot for Scaled Predictions

- Residuals:
  - Residuals show Predicted values are mostly normal aside from a few outliers
- QQ Plot:
  - Shows normal distribution with most residuals within 3 standard deviations



Conclusion

Conclusions

- Which factors are the most important predictors of home price?
  - Principal Components 1 and 4 have the highest standardized betas (0.696 and 0.369, respectively)
    - Principal 1 is composed mostly of the following variables: **sqft_living, bathrooms, grade**
    - Principal 4 is composed mostly of the following variables: **zipcode_cluster and sqft_lot**
- R^2 overall explains 79% of the variation in home prices
- Significant outliers in the field, some unique homes may need to be priced individually by experts in the field

Thank You!
Questions?

Appendix: Test and Training Data

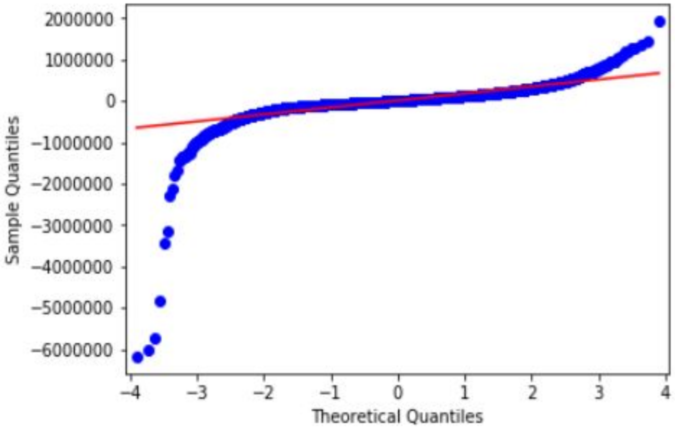
Cross Validation

- Split data into 75% training and 25% test (randomly sampled)
- Ran initial regression on training data and removed influential observations
- Reran regression without influential points
- Applied regression equation to test data to get predictions
- Scaled predictions back to original range

Comparison of Accuracy on Full Data & Test Data

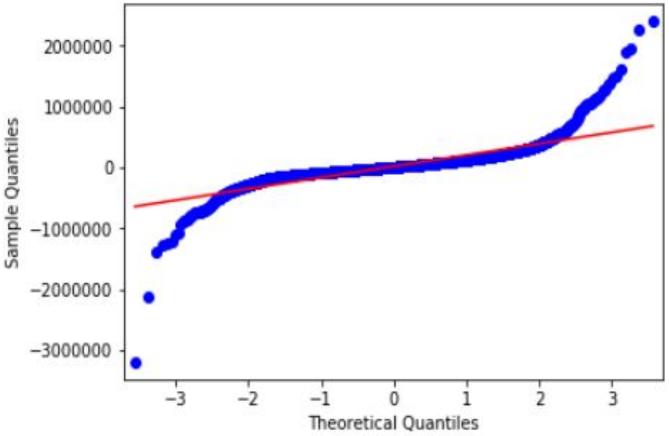
Accuracy Full Dataset (21,613 rows)

- R^2 = 0.76
- RMSE = 168,778
- Mean Absolute Error = 96.050
- Median Absolute Error = 65,520



Accuracy for Test Dataset (5,403 rows)

- R^2=0.738
- RMSE=186,507
- Mean Absolute Error = 105,534
- Median Absolute Error = 67,683



- The disparity in results shows the model slightly overfit the training data
- Possible reason for this is that we removed influential points from training data, but not the test data